(19) RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL
DE LA PROPRIÉTÉ INDUSTRIELLE

PARIS

(11) N° de publication : 2 825 168
(à n'utiliser que pour les commandes de reproduction)

(21) N° d'enregistrement national : 01 07006

(51) Int Cl[7] : G 06 F 17/10 // G 06 F 17/30

(12) DEMANDE DE BREVET D'INVENTION A1

(22) Date de dépôt : 23.05.01.

(30) Priorité :

(43) Date de mise à la disposition du public de la demande : 29.11.02 Bulletin 02/48.

(56) Liste des documents cités dans le rapport de recherche préliminaire : *Se reporter à la fin du présent fascicule*

(60) Références à d'autres documents nationaux apparentés :

(71) Demandeur(s) : *FRANCE TELECOM Société anonyme* — FR.

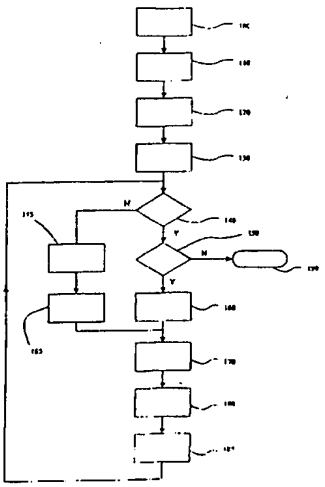
(72) Inventeur(s) : BOULLE MARC.

(73) Titulaire(s) :

(74) Mandataire(s) : CABINET LE GUEN ET MAILLET.

(54) PROCEDE DE DISCRETISATION D'ATTRIBUTS D'UNE BASE DE DONNEES.

(57) Méthode de discrétisation d'un attribut d'une base de données contenant une population d'individus, ledit attribut, dit attribut source, pouvant prendre plusieurs modalités, la méthode étant caractérisée en ce que, dans une première étape, on regroupe lesdites modalités de l'attribut source en groupes élémentaires et, à partir du tableau de contingence de l'attribut source et d'un attribut cible, on détermine, dans une seconde étape, parmi un ensemble de couples de groupes élémentaires, le couple de groupes élémentaires dont la fusion diminue le plus fortement la probabilité d'indépendance de l'attribut source et de l'attribut cible, et que l'on fusionne dans une troisième étape le couple de groupes élémentaires ainsi déterminé, lesdites seconde et troisième étapes étant itérées tant qu'il existe un couple de groupes élémentaires permettant de diminuer ladite probabilité d'indépendance.



FR 2 825 168 - A1

La présente invention concerne une méthode de discrétisation d'attributs d'une base de données. L'invention trouve particulièrement application dans l'exploitation statistique des données, notamment dans le domaine de l'apprentissage supervisé.

L'analyse statistique des données (encore appelée «data mining ») a pris un essor considérable ces dernières années avec l'extension du commerce électronique et l'apparition de très grandes bases de données. Le data mining vise de manière générale à explorer, classifier et extraire des règles d'associations sous-jacentes au sein d'une base de données. Il est notamment utilisé pour construire des modèles de classification ou de prédiction. La classification permet d'identifier au sein de la base de données des catégories à partir de combinaisons d'attributs, puis de ranger les données en fonction de ces catégories. Par exemple, si la base de données est relative à des achats de produits par des consommateurs, ceux-ci pourront être rangés en différentes catégories : clients fidèles, clients occasionnels, clients recherchant les produits soldés, clients recherchant les produits haut de gamme etc. La prédiction, quant à elle, vise à décrire comment un ou plusieurs attributs de la base de données se comporteront dans le futur. Dans l'exemple de la base de données d'achats évoqué plus haut, il pourra être intéressant de prévoir le comportement de ces consommateurs en fonction d'une baisse ou d'une hausse de prix de tel ou tel produit.

Un des objectifs du data mining dit « supervisé » est la construction d'un modèle prédictif visant à prédire un attribut déterminé. Cette construction consiste à chercher parmi les attributs de la base de données considérée à identifier celui ou ceux qui présentent la plus forte dépendance statistique avec un attribut cible et à décrire cette dépendance. Par exemple, si l'on a classé les consommateurs en fonction de leurs montants d'achats annuels en différentes catégories de consommation: grosse consommation, moyenne consommation, faible consommation, il sera intéressant de déterminer quels sont les attributs de la base de données achats qui sont les plus corrélés (ou de manière équivalente, les moins indépendants statistiquement) de l'attibut donnant la classe de consommation. On notera qu'au lieu de d'attribut cible « catégorie de consommation », on aurait pu prendre directement l'attribut « montant d'achats annuels ».

De manière générale, les valeurs (encore appelées modalités) prises par un attribut peuvent être numériques (par exemple un montant d'achats) ou symbolique (par exemple une catégorie de consommation). On parle dans le premier cas d'attribut numérique et dans le second cas d'attribut symbolique.

Certaines méthodes de data mining supervisé requièrent une « discrétisation » des attributs numériques. On entend ici par discrétisation d'un attribut numérique un découpage du domaine des valeurs prises par un attribut en un nombre fini d'intervalles. Si le domaine en question est une plage de valeurs continues la discrétisation se traduira par une quantification de cette plage. Si ce domaine est déjà constitué de valeurs discrètes ordonnées, la discrétisation aura pour fonction de regrouper ces valeurs en groupes de valeurs consécutives.

La discrétisation des attributs numériques a été largement traitée dans la littérature. On en trouvera par exemple une description dans l'ouvrage de Zighed et al. intitulé « Graphes d'induction » publié chez HERMES Science Publications. On distingue deux types de méthodes de discrétisation : les méthodes descendantes et les méthodes ascendantes. Les méthodes descendantes partent de l'intervalle complet à discrétiser et cherche le meilleur point de coupure de l'intervalle en optimisant un critère prédéterminé. Les méthodes ascendantes partent d'intervalles élémentaires et cherchent la meilleure fusion de deux intervalles adjacents en optimisant un critère prédéterminé. Dans les deux cas, elles sont appliquées itérativement jusqu'à ce qu'un critère d'arrêt soit satisfait.

Une méthode de discrétisation ascendante utilisant le critère du χ^2 est connu dans la littérature sous le nom de ChiMerge. De même une méthode de discrétisation descendante utilisant le critère du χ^2 est connu sous le nom de ChiSplit.

Avant de présenter la méthode ChiMerge on rappellera tout d'abord que le critère du χ^2 permet sous certaines hypothèses de déterminer le degré d'indépendance de deux variables aléatoires. Soit S un attribut source et T un attribut cible. On supposera pour fixer les idées que S présente quatre modalités a,b,c,d et T trois modalités A,B,C. Le Tableau 1 montre le tableau de contingence des variables S et T avec les conventions suivantes :

n_{ij} est le nombre d'individus observés pour la $i^{ème}$ modalité de la variable S et la $j^{ème}$ modalité de la variable T . n_{ij} est encore appelé effectif observé de la case (i,j) ;

$n_{i.}$ est le nombre total d'individus pour la $i^{ème}$ modalité de la variable S . $n_{i.}$ est encore appelé effectif observé de la ligne i ;

$n_{.j}$ est le nombre total d'individus pour la $j^{ème}$ modalité de la variable T . $n_{.j}$ est encore appelé effectif observé de la colonne j ;

N est le nombre total d'individus.

S/T	A	B	C	Total
A	n_{11}	n_{12}	n_{13}	$n_{1.}$
B	n_{21}	n_{22}	n_{23}	$n_{2.}$
C	n_{31}	n_{32}	n_{33}	$n_{3.}$
D	n_{41}	n_{42}	n_{43}	$n_{4.}$
E	n_{51}	n_{52}	n_{53}	$n_{5.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

Tableau 1

De manière générale, on notera I et J respectivement le nombre de modalités de l'attribut S et le nombre de modalités de l'attribut T.

On définit l'effectif théorique e_{ij} de la case (i,j) par $e_{ij}=\frac{n_{i.}n_{.j}}{N}$. e_{ij} représente le nombre d'individus qui serait observé dans la case du tableau de contingence dans le cas de variables indépendantes. L'écart à l'indépendance des variables S et T est mesuré par :

$$\chi^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} \qquad (1)$$

Plus la valeur de χ^2 est élevée, moins l'hypothèse d'indépendance des variables aléatoires S et T est probable. On parle par abus de langage de probabilité d'indépendance des variables.

Plus précisément χ^2 est une variable aléatoire dont on peut montrer que la densité suit une loi dite du χ^2 à $(I\text{-}1).(J\text{-}1)$ degrés de liberté. La loi du χ^2 est celle suivie par une somme quadratique de valeurs aléatoires normales centrées. Elle a de fait l'expression d'une loi γ et tend vers une loi gaussienne lorsque le nombre de degrés de liberté est élevé.

Par exemple si I=5 et J=3, le nombre de degrés de liberté vaut 8. Si la valeur de χ^2 calculée par (1) vaut 20, la loi du χ^2 à 8 degrés de liberté donne une probabilité d'indépendance de S et T de 1%.

Nous présenterons ci-après la méthode de discrétisation ChiMerge. Nous nous plaçons dans le cas général d'un attribut source S à I modalités et d'un attribut T à J modalités. La méthode ChiMerge considère seulement deux lignes consécutives i et i+1 du tableau de contingence. Soit $q'_1,q'_2,..,q'_J$ la distribution locale (c'est-à-dire dans le contexte local des lignes consécutives i et i+1) de probabilité des modalités pour l'attribut cible T. Si $n_{i.}$ est l'effectif de la ligne i et $n_{i+1.}$ est l'effectif de la ligne i+1, les effectifs observés et théoriques de la ligne i s'expriment respectivement par $n_{ij}=a_{ij}n_{i.}$ et $e_{ij}=q'_j n_{i.}$ où les a_{ij} représentent les proportions d'effectifs observés pour la ligne i. De même, les effectifs observés et théoriques de la ligne i+1 s'expriment respectivement par $n_{i+1,j}=a_{i+1,j}n_{i+1,.}$ et $e_{i+1,j}=q'_j n_{i+1,.}$ où les $a_{i+1,j}$ représentent les proportions observées de modalités de T pour la ligne i+1. La distribution locale de probabilité $q'_1,q'_2,..,q'_J$ des modalités de l'attribut cible peut être exprimée par :

$$q'_j=\frac{a_{ij}n_{i.}+a_{i+1,j}n_{i+1,.}}{n_{i.}+n_{i+1,.}} \qquad (2)$$

Selon la méthode ChiMerge, on calcule la valeur du χ^2 pour les lignes i et $i+1$, soit, en tenant compte du fait que $\sum_{j=1}^{J} q'_j = \sum_{j=1}^{J} a_{ij} = 1$:

$$\chi^2_{i,i+1} = n_{i.}\left(\sum_{j=1}^{J}\frac{a_{ij}^2}{q'_j} - 1\right) + n_{i+1,.}\left(\sum_{j=1}^{J}\frac{a_{i+1,j}^2}{q'_j} - 1\right) \quad (3)$$

soit encore après transformation :

$$\chi^2_{i,i+1} = \frac{n_{i.}n_{i+1,.}}{n_{i.}+n_{i+1,.}}\sum_{j=1}^{J}\frac{\left(a_{ij}-a_{i+1,j}\right)^2}{q'_j} \quad (4)$$

$\chi^2_{i,i+1}$ est une variable aléatoire suivant une loi du χ^2 à J -1 degrés de liberté. La méthode ChiMerge propose de fusionner les lignes i et $i+1$ si :

$$prob(\chi^2_{i,i+1}, J-1) \leq p_{Th} \quad (5)$$

où $prob(\alpha,K)$ désigne la probabilité que $\chi^2 \geq \alpha$ pour la loi du χ^2 à K degrés de libertés et p_{Th} est une valeur de seuil prédéterminée paramétrant la méthode. En pratique, la valeur $prob(\alpha,K)$ est obtenue à partir d'une table classique du χ^2 donnant la valeur de α en fonction de $prob(\alpha,K)$ et de K.

La condition (5) exprime que la probabilité d'indépendance de S et T au vu des deux lignes considérées est inférieure à une valeur de seuil. La fusion de lignes consécutives est itérée tant que la condition (5) est vérifiée. La fusion de deux lignes entraîne le regroupement de leurs modalités et la sommation de leurs effectifs. Par exemple dans le cas d'un attribut numérique à valeurs continues on a avant fusion :

$[s_i, s_{i+1}[$	$n_{i,1}$	$n_{i+1,2}$	....	$n_{i,J}$	$n_{i,.}$

$[s_{i+1},s_{i+2}[$	$n_{i+1,1}$	$n_{i+1,2}$	....	$n_{i+1,J}$	$n_{i+1,.}$

Tableau 2

et après fusion :

$[s_i,s_{i+2}[$	$n_{i,1}+n_{i+1,1}$	$n_{i+1,2}+n_{i+1,2}$	....	$n_{i,J}+n_{i+1,J}$	$n_{i,.}+n_{i+1,.}$

Tableau 3

Un premier problème soulevé par l'emploi de la méthode ChiMerge est le choix du paramètre p_{Th} qui ne doit pas trop élevé sous peine de fusionner toutes les lignes ni trop faible sous peine de n'en fusionner aucune paire. En pratique, il est très difficile de trouver un compromis.

Un second problème intrinsèque à cette méthode est d'opérer localement sans tenir compte de l'ensemble des modalités (ou du nombre d'intervalles) de l'attribut source. On ne sait pas *a priori* si le résultat de la discrétisation est globalement optimal sur cet ensemble.

En outre, la méthode ChiMerge est limitée à une discrétisation mono-dimensionnelle en ce sens qu'elle ne peut opérer que sur un seul attribut source à la fois et non sur un p-uplet d'attributs.

Enfin, la méthode ChiMerge ne permet pas de mesurer la probabilité d'indépendance entre un attribut source et un attribut cible et, par voie de conséquence, pour un attribut cible donné, de classer des attributs source en fonction de leurs probabilités d'indépendance vis à vis de l'attribut cible.

L'objectif de la présente invention est de proposer une méthode de discrétisation d'attributs qui ne présente pas les inconvénients et limitations énoncés ci-dessus. A cet effet, l'invention est définie par une méthode de discrétisation d'un attribut d'une base de données contenant une population d'individus, ledit attribut, dit attribut source, pouvant prendre plusieurs modalités, ladite méthode comprenant une première étape dans laquelle on regroupe lesdites modalités de l'attribut source en groupes

élémentaires et, une seconde étape dans laquelle on détermine, à partir du tableau de contingence de l'attribut source et d'un attribut cible, parmi un ensemble de couples de groupes élémentaires, le couple de groupes élémentaires dont la fusion diminue le plus fortement la probabilité d'indépendance de l'attribut source et de l'attribut cible, et une troisième étape dans laquelle on fusionne le couple de groupes élémentaires ainsi déterminé, lesdites seconde et troisième étapes étant itérées tant qu'il existe un couple de groupes élémentaires permettant de diminuer ladite probabilité d'indépendance.

Afin de déterminer le couple de groupes élémentaires dans la seconde étape, on pourra estimer pour chaque couple de groupes élémentaires dudit ensemble, la valeur du χ^2 du tableau de contingence après fusion dudit couple et l'on sélectionnera le couple produisant la valeur du χ^2 après fusion la plus élevée.

Avantageusement, pour chaque couple de groupes élémentaires, on calcule la variation du χ^2 du tableau de contingence avant et après fusion dudit couple. Les variations du χ^2 associées aux différents couples seront alors triées sous forme de liste de valeurs décroissantes et que l'on sélectionnera le premier couple de la liste.

Le couple de groupes élémentaires étant sélectionné, on procèdera à la fusion dudit couple si la probabilité du χ^2 relative au tableau de contingence après fusion dudit couple est inférieure à la probabilité du χ^2 relative au tableau de contingence avant fusion.

Selon une variante, les probabilités du χ^2 relatives au tableau de contingence avant et après fusion sont exprimées de manière logarithmique.

Typiquement, ledit ensemble de couples de groupes élémentaires est constitué de tous les couples de groupes voisins au sens d'une relation de voisinage prédéterminée.

On recherche de préférence parmi les couples de groupes élémentaires voisins ceux comprenant au moins un groupe présentant au moins un effectif théorique par case du tableau de contingence inférieur à un effectif minimum prédéterminé et on les identifie comme couples prioritaires au moyen d'une information d'identification.

Dans ce cas, s'il existe un ou des couples prioritaires, on fusionne le couple prioritaire produisant la valeur du χ^2 après fusion la plus élevée.

Selon un premier mode de réalisation, l'attribut source étant un attribut numérique mono-dimensionnel, les groupes élémentaires voisins sont constitués par des intervalles adjacents.

Selon un second mode de réalisation, l'attribut source étant un attribut numérique multi-dimensionnel formé par pluralité d'attributs numériques mono-dimensionnels et les individus de la population étant représentés par des points dans l'espace desdits attributs, lesdits groupes élémentaires sont les cellules de Voronoï de cet espace, contenant lesdits points.

Dans ce cas, on construit le graphe de Delaunay associé aux cellules de Voronoï et l'on élimine de ce graphe tout arc joignant deux cellules voisines en passant par une troisième, les couples de groupes élémentaires voisins étant alors donnés par les arcs du graphe de Delaunay après l'étape d'élimination.

Selon un troisième mode de réalisation, l'attribut source est de type symbolique.

L'invention concerne encore une méthode d'évaluation de la dépendance d'un attribut numérique bi-dimensionnel, formé par un couple d'attributs numériques mono-dimensionnels, vis à vis d'un attribut cible. Les individus de la population sont représentés par des points dans le plan desdits attributs. Selon cette méthode, on discrétise l'attribut bi-dimensionnel par la méthode de discrétisation multi-dimensionnelle mentionnée plus haut et l'on visualise par des moyens de visualisation des goupes de cellules de Voronoï fusionnées par ladite méthode.

L'invention concerne enfin un logiciel de data mining comprenant un programme de discrétisation d'au moins un attribut d'une base de données, tel que son exécution sur un ordinateur effectue les étapes de la méthode exposée ci-dessus.

Les caractéristiques de l'invention mentionnées ci-dessus, ainsi que d'autres, apparaîtront plus clairement à la lecture de la description suivante d'un exemple de réalisation, ladite description étant faite en relation avec les dessins joints, parmi lesquels :

la Fig. 1 illustre sous forme d'organigramme la méthode de discrétisation d'attributs selon un mode de réalisation de l'invention ;

la Fig. 2 illustre un premier exemple de discrétisation d'un attribut symbolique;

la Fig. 3 illustre un second exemple de discrétisation d'un attribut symbolique avant et après fusion;

la Fig. 4 représente un exemple de diagramme de Voronoï ;

la Fig. 5 représente le diagramme de Delaunay associé au diagramme de Voronoï de la Fig. 4 ;

la Fig. 6 représente un ensemble d'individus projetés sur le plan de deux attributs numériques ;

la Fig. 7 représente le diagramme de Delaunay associé à l'ensemble d'individus de la Fig. 6 ;

la Fig. 8 représente les zones de discrétisation associées à l'ensemble d'individus de la Fig. 5.

Une première idée générale à la base de l'invention est de discrétiser un attribut source en optimisant un critère statistique portant sur l'ensemble du tableau de contingence. Une seconde idée générale à la base de l'invention est d'extrapoler cette discrétisation au cas multi-dimensionnel en faisant appel à un graphe de Delaunay.

Nous exposerons l'invention tout d'abord dans le cas d'un attribut S numérique mono-dimensionnel à valeurs continues. Après avoir ordonné les modalités de S, l'ensemble de ces modalités peut être découpé en intervalles élémentaires $S_i=[s_i,s_{i+1}[$, $i=1,..,I$. Nous souhaitons évaluer le degré d'indépendance de cet attribut avec un attribut cible T de modalités T_j, $j=1,..,J$. Ces modalités T_j peuvent être des modalités symboliques ou numériques. Dans ce dernier cas elles peuvent être des valeurs discrètes ou des intervalles de valeurs continues. On peut représenter le tableau de contingence :

S/T	T_1	T_2	...	T_J	Total
S_1	$n_{1,1}$	$n_{1,2}$	...	$n_{1,J}$	$n_{1,.}$
⋮	⋮	⋮	⋮	⋮	⋮
S_i	$n_{i,1}$	$n_{i,2}$	...	$n_{i,J}$	$n_{i,.}$
S_{i+1}	$n_{i+1,1}$	$n_{i+1,2}$	...	$n_{i+1,J}$	$n_{i+1,.}$
⋮	⋮	⋮	⋮	⋮	⋮

S_I	$n_{I,1}$	$n_{I,2}$	...	$n_{I,J}$	$n_{I,.}$
Total	$n_{.,1}$	$n_{.,2}$	...	$n_{.,J}$	N

Tableau 4

D'après (1) la valeur du χ^2 sur l'ensemble du tableau peut s'exprimer par :

$$\chi^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} \qquad (6)$$

Soit encore en notant $q_1,q_2,..,q_J$ la distribution de probabilité des modalités de l'attribut cible et a_{ij} les proportions d'effectifs observés pour la ligne i et en remarquant que $e_{ij}=q_j n_{i,.}$, $n_{ij}=a_{ij}n_{i,.}$ et $\sum_{j=1}^{J}q_j=\sum_{j=1}^{J}a_{ij}=1$:

$$\chi^2=\sum_{i=1}^{I}n_{i,.}\sum_{j=1}^{J}\left(\frac{a_{ij}^{\ 2}}{q_j}-1\right)=\sum_{i=1}^{I}\chi^2_{(i)} \qquad (7)$$

où $\chi^2_{(i)}$ est la valeur du χ^2 pour la ligne i. L'expression (7) signifie que le χ^2 est additif par rapport aux lignes du tableau.

Supposons maintenant que deux lignes consécutives i et $i+1$ soient fusionnées. La valeur du χ^2 après fusion, notée $\chi^2_{f(i,i+1)}$ peut s'écrire :

$$\chi^2_{f(i,i+1)}=\sum_{k<i}\chi^2_{(k)}+\chi^2_{(i,i+1)}+\sum_{k>i+1}\chi^2_{(k)} \qquad (8)$$

où $\chi^2_{(i,i+1)}$ est la valeur du χ^2 pour la ligne résultant de la fusion, c'est-à-dire :

$$\chi^2_{(i,i+1)}=(n_{i,.}+n_{i+1,.})\sum_{j=1}^{J}\left(\frac{a'^{\,2}_{ij}}{q_j}-1\right) \quad \text{avec} \quad a'_{ij}=\frac{n_{ij}+n_{i+1,j}}{n_{i,.}+n_{i+1,.}} \qquad (9)$$

L'expression (8) peut s'exprimer simplement en fonction de valeur du χ^2 avant fusion :

$$\chi^2_{f(i,i+1)}=\chi^2+\chi^2_{(i,i+1)}-\chi^2_{(i)}-\chi^2_{(i+1)}=\chi^2+\Delta\chi^2_{(i,i+1)} \quad (10)$$

où $\Delta\chi^2_{(i,i+1)}$ est la variation du χ^2 résultant de la fusion des lignes i et i+1. La valeur de $\Delta\chi^2_{(i,i+1)}$ peut être calculée explicitement en fonction des proportions d'effectifs des lignes i et i+1 :

$$\Delta\chi^2_{(i,i+1)}=\left(\frac{n_{i,.}+n_{i+1,.}}{n_{i,.}n_{i+1,.}}\right)\sum_{j=1}^{J}\frac{\left(a_{ij}-a_{i+1,j}\right)^2}{q_j} \quad (11)$$

La liste des valeurs de $\Delta\chi^2_{(i,i+1)}$ est triée par valeurs décroissantes. Soit $\Delta\chi^2_{(i_0,i_0+1)}$ premier élément de la liste. On teste alors si :

$$prob\left(\chi^2_{f(i_0,i_0+1)},(I-2)(J-1)\right)\leq prob\left(\chi^2,(I-1)(J-1)\right) \quad (12)$$

On notera que la loi du χ^2 pour le premier terme n'a plus que (I-2)(J-1) degrés de liberté suite à la fusion. En pratique, étant donné les faibles valeurs que peuvent prendre les termes de (12), la comparaison portera avantageusement sur les logarithmes de ces probabilités.

La condition (12) traduit une diminution de la probabilité d'indépendance de S et T après fusion des lignes i_0 et i_0+1. Etant donné la valeur négative $\Delta\chi^2_{(i_0,i_0+1)}$, la valeur du χ^2 ne peut que décroître avec la fusion. Etant donné que $prob(\alpha,K)$ est une fonction décroissante de α et croissante de K, la relation (12) ne peut être vérifiée que grâce à la diminution du nombre de degrés de liberté. La diminution de la probabilité d'indépendance sera d'autant plus importante que $\Delta\chi^2_{(i_0,i_0+1)}$ sera faible en valeur absolue, c'est à dire d'après la relation (11) que les proportions observées pour les lignes considérées seront plus proches et ce pour les proportions q_j les plus faibles.

Si la condition (12) est vérifiée, on fusionne les lignes i_0 et i_0+1. En revanche, si la condition (12) n'est pas vérifiée, alors elle n'est vérifiée pour aucun indice i par suite de la décroissance de $prob(\alpha,K)$ en fonction de α. Le processus de fusion est alors arrêté.

Si les lignes i_0 et i_0+1 ont été fusionnées, on met à jour la liste des valeurs $\Delta\chi^2_{(i,i+1)}$. On notera que cette mise à jour ne concerne en fait que les valeurs relatives

aux lignes contiguës aux lignes fusionnées à savoir les lignes d'indices i_0-1 et i_0+2 avant fusion (si elles existent). Le processus de fusion est itéré tant que la condition (12) est satisfaite.

La méthode décrite ci-dessus conduit à une discrétisation *ad hoc* du domaine des modalités, c'est-à-dire à une discrétisation qui minimise l'indépendance entre l'attribut source et l'attribut cible sur l'ensemble du domaine. La méthode de discrétisation permet de regrouper des intervalles adjacents ayant des comportements de prédiction similaires vis à vis de l'attribut cible, le regroupement étant arrêté lorsqu'il nuit à la qualité de prédiction, en d'autres termes lorsqu'il ne fait plus décroître la probabilité d'indépendance des attributs.

On obtient par fusions successives un tableau de contingence dont le nombre de lignes se réduit et dont les effectifs par case augmentent. Afin de pouvoir tirer des conclusions fiables quant à la dépendance ou l'indépendance des attributs source et cible il est souhaitable d'avoir un effectif minimum par case. Il est communément admis que le test du χ^2 est fiable pour des effectifs théoriques supérieurs à 5 par case. Qui plus est, une distribution inhomogène étant plus probable pour une faible population que pour une population plus importante, on observe pour de faibles valeurs d'effectifs théoriques e_{ij} un phénomène, dit de « sur-apprentissage » dans lequel, à partir d'une valeur élevée du χ^2 on conclut indûment à une dépendance des attributs. On pourra alors convenir de respecter un effectif théorique minimum par case. On peut montrer qu'un effectif moyen minimum de l'ordre de $\log_2(10N)$ (où N est le nombre total d'individus) par case permet d'éviter de conclure de manière erronée à la dépendance des attributs. La méthode de discrétisation est alors adaptée de la manière suivante : on accorde d'abord la priorité aux fusions de lignes vérifiant (12) qui permettent de vérifier un critère d'effectif minimum. Le critère d'effectif minimum pourra, par exemple, s'écrire pour la ligne i_0:

$$e_{i_0,j} \geq \log_2(10N), j=1,..,J \qquad (13)$$

Pour ce faire, on pourra marquer d'un drapeau les couples de lignes dont au moins l'une d'elles ne vérifie pas la condition d'effectif minimum (13) et l'on fusionnera le premier couple de lignes d'indices i_0 et i_0+1 portant un tel drapeau. Après fusion on met à jour les drapeaux des lignes adjacentes i_0-1 et i_0+2 en fonction

de l'effectif atteint par la ligne fusionnée. Lorsque toutes les lignes ont atteint l'effectif minimum, seule la condition (12) est prise en compte puisque critère le critère d'effectif minimum est rempli.

La Fig. 1 illustre l'algorithme d'un exemple de méthode de discrétisation selon l'invention.

L'algorithme débute par une étape 100 de partition du domaine de valeurs de la loi source en intervalles élémentaires ordonnés. La valeur de χ^2 pour le tableau de contingence et les valeurs $\chi^2_{(i)}$ pour les I lignes du tableau sont calculées en 110. Les valeurs $\Delta\chi^2_{(i,i+1)}$ sont ensuite déduites des valeurs $\chi^2_{(i)}$ à l'étape 120 et triées par valeurs décroissantes sous forme de liste en 130. Chaque élément de la liste correspond à la fusion possible d'un couple de lignes i et i+1. L'étape 140 teste si la condition d'effectif minimum (13) est vérifiée. Dans l'affirmative, on passe directement au test 150. Dans la négative, on poursuit par l'étape 145.

A l'étape 145, on donne priorité (au moyen de drapeaux) aux couples de lignes dont l'une d'entre elles au moins n'a pas atteint l'effectif minimum et l'on sélectionne en 165 le premier couple prioritaire de la liste que nous noterons (i_0, i_0+1). Le processus se poursuit en 170.

A l'étape 150, on teste si le premier élément de la liste vérifie la condition (12). Si ce n'est pas le cas, le processus se termine en 190. En revanche, dans l'affirmative, on sélectionne en 160 le premier couple de la liste, que nous noterons également (i_0 , i_0+1) et l'on poursuit par l'étape 170.

A l'étape 170, les lignes i_0 et i_0+1 du couple sélectionné sont fusionnées, c'est-à-dire les intervalles S_i et S_{i+1} sont concaténés. La nouvelle valeur de $\chi^2_{(i_0)}$ est ensuite calculée en 180 ainsi que les nouvelles valeurs de $\Delta\chi^2_{(i_0-1,i_0)}$ et $\Delta\chi^2_{(i_0,i_0+1)}$ pour les intervalles adjacents, s'ils existent. En 185, La liste des valeurs $\Delta\chi^2_{(i,i+1)}$ est mise à jour: les anciennes valeurs $\Delta\chi^2_{(i_0-1,i_0)}$ et $\Delta\chi^2_{(i_0,i_0+1)}$ sont supprimées et les nouvelles valeurs sont stockées. La liste des valeurs $\Delta\chi^2_{(i,i+1)}$ est avantageusement organisée sous forme d'arbre binaire de recherche équilibré permettant de gérer les insertions/suppressions tout en maintenant la relation d'ordre dans la liste. Ainsi, il n'est pas nécessaire de trier complètement la liste à chaque étape. La liste des drapeaux est également mise à jour. Après la mise à jour, le processus retourne à l'étape de test 140.

Selon une variante de réalisation, la liste est constituée par les valeurs (positives) $\chi^2_{f(i,i+1)}$ au lieu d'être constituée par valeurs (négatives) $\Delta\chi^2_{(i,i+1)}$.

Au terme du processus de discrétisation, on dispose de la valeur du χ^2 de l'attribut discrétisé. Ainsi, si l'on procède à la discrétisation d'une pluralité d'attributs source S_k, on peut comparer leur capacité prédictive vis à vis de l'attribut cible en comparant les probabilités $prob(\chi_k^2, \alpha_k)$ où les χ_k^2 et α_k sont les valeurs de χ^2 et les degrés de liberté respectifs des attributs discrétisés.

– Nous avons supposé jusqu'à présent que l'attribut S était numérique mono-dimensionnel à valeurs continues. La méthode de discrétisation exposée ci-dessus est encore applicable lorsque S est à valeurs numériques discrètes. Les modalités numériques sont d'abord ordonnées pour former les lignes du tableau de contingence de S et T puis regroupées par groupes élémentaires, un groupe élémentaire pouvant, le cas échéant, ne contenir qu'un seul élément. La méthode de discrétisation opère selon le même principe que précédemment, en fusionnant les groupes élémentaires tant que la probabilité d'indépendance de S et T diminue.

La méthode de discrétisation peut encore opérer sur des attributs symboliques, à la différence qu'il n'existe pas nécéssairement de relation d'ordre total entre les modalités de l'attribut. Si une telle relation d'ordre existe, on peut se ramener au cas précédent en ordonnant les modalités selon cette relation d'ordre. La Fig. 2 illustre cette situation : les individus sont regroupés par groupes élémentaires $G_1, G_2, .., G_I$, chaque groupe contenant les individus relatifs à une modalité ou à un intervalle de modalités (au sens de la relation d'ordre précitée). Les groupes sont équivalents aux lignes du tableau de contingence. Ils peuvent être ordonnés au sein d'un graphe linéaire, chaque noeud correspondant à un groupe. La fusion ne peut être réalisée que selon les arcs de ce graphe, entre groupes voisins. En revanche, si l'ensemble des modalités de l'attribut source n'est pas pourvue d'une relation d'ordre total, on peut néanmoins définir des relations de voisinage par des arcs d'un graphe, comme représenté dans la partie gauche de la Fig. 3. Les arcs indiquent les fusions possibles entre les groupes. Après fusion de deux groupes, les arcs du graphe sont réorganisés. La partie droite de la Fig. 3 représente une réorganisation du graphe après fusion des groupes 3 et 4. La méthode de discrétisation opère ici sur les noeuds du graphe de la même façon qu'elle opérait précédemment sur les lignes du tableau de contigence.

Le fonctionnement de la méthode de discrétisation sera illustré à l'aide d'un exemple relatif à une base de données contenant des attributs de fleurs de la famille des Iris. La population de la base de données considérée est de 150 individus. Nous envisagerons l'attribut source « largeur de sépale » et l'attribut cible classe de la

fleur : Iris setosa, Iris versicolor, Iris virginica. Dans cet exemple, l'attribut source est un attribut numérique à valeurs continues et l'attribut cible est un attribut symbolique à 3 modalités. Le tableau de contingence est donné ci-après :

Largeur de sépale	Iris versicolor	Iris virginica	Iris setosa	Total
2	1	0	0	1
2,2	2	1	0	3
2,3	3	0	1	4
2,4	3	0	0	3
2,5	4	4	0	8
2,6	3	2	0	5
2,7	5	4	0	9
2,8	6	8	0	14
2,9	7	2	1	10
3	8	12	6	26
3,1	3	4	5	12
3,2	3	5	5	13
3,3	1	3	2	6
3,4	1	2	9	12
3,5	0	0	6	6
3,6	0	1	2	3
3,7	0	0	3	3
3,8	0	2	4	6
3,9	0	0	2	2
4	0	0	1	1
4,1	0	0	1	1
4,2	0	0	1	1
4,4	0	0	1	1
Total	50	50	50	150

Tableau 5

Lors de l'initialisation, on partitionne le domaine des modalités de la largeur de sépale $[0,+\infty[$ en 23 intervalles élémentaires : $]-\infty; 2,1]$, $]2,1; 2,25]$ ... $]4,15; 4,3]$, $]4,3; +\infty[$. La valeur du χ^2 est de 88,36. En prenant la loi du χ^2 à 44 degrés de libertés correspondante (44=(23-1)*(3-1)), on obtient une probabilité d'indépendance de $8,3\ 10^{-5}$. Comme indiqué dans le tableau 6, on calcule alors le χ^2 résultant de chaque

fusion d'intervalles : $\chi^2_{f(i,i+1)}$. Par exemple, la fusion des intervalles]-∞; 2,1],]2,1; 2,25] donne un nouvel intervalle]-∞; 2,25] et le χ^2 résultant de la nouvelle table réduite a une valeur de 87,86.

Intervalle fusionné	$\chi^2_{f(i,i+1)}$
]∞ ; 2,25]	87,86
]2,10; 2,35]	87,44
]2,25; 2,45]	87,72
]2,35; 2,55]	85,09
]2,45; 2,65]	88,18
]2,55; 2,75]	88,33
]2,65; 2,85]	87,83
]2,75; 2,95]	84,49
]2,85; 3,05]	83,18
]2,95; 3,15]	87,03
]3,05; 3,25]	88,29
]3,15; 3,35]	88,12
]3,25; 3,45]	84,86
]3,35; 3,55]	87,20
]3,45; 3,65]	87,03
]3,55; 3,75]	87,36
]3,65; 3,85]	87,03
]3,75; 3,95]	87,36
]3,85; 4,05]	88,36
]3,95; 4,15]	88,36
]4,05; 4,25]	88,36
]4,15 ; +∞ [	88,36

Tableau 6

On cherche alors la fusion qui maximise le χ^2 Ici, la valeur maximale du χ^2 résultant d'une fusion est de 88,36, atteinte par exemple pour la fusion des deux derniers intervalles]4,15; 4,3] et]4,3; +∞[. En prenant la loi du χ^2 à 42 degrés de liberté correspondante (il y a un intervalle en moins), on obtient une probabilité d'indépendance de $3{,}8\ 10^{-5}$. La probabilité d'indépendance diminuant, la discrétisation est améliorée et on réalise la fusion correspondante. On recommence ces étapes tant qu'il y a amélioration de la discrétisation. Le tableau 7 illustre les étapes successives

de discrétisation. Les chiffres en gras indiquent que l'effectif minimum est atteint, au sens de la relation (13). Ici, étant donné que les modalités de l'attribut cible sont équiréparties (q_1=q_2=q_3) la relation (13) est équivalente à un effectif théorique par ligne de 33 ($3.\log_2(10*150)$). Lorsque cet effectif est atteint pour toutes les lignes, on ne tient plus compte du critère d'effectif minimum.

Largeur de sépale	Iris versicolor	Iris virginica	Iris	Total
2	1	0	0	1
2,2	2	1	0	3
2,3	3	0	1	4
2,4	3	0	0	3
2,5	4	4	0	8
2,6	3	2	0	5
2,7	5	4	0	9
2,8	6	8	0	14
2,9	7	2	1	10
3	8	12	6	26
3,1	3	4	5	12
3,2	3	5	5	13
3,3	1	3	2	6
3,4	1	2	9	12
3,5	0	0	6	6
3,6	0	1	2	3
3,7	0	0	3	3
3,8	0	2	4	6
3,9	0	0	2	2
4	0	0	1	1
4,1	0	0	1	1
4,2	0	0	1	1
4,4	0	0	1	1
Total	50	50	50	150

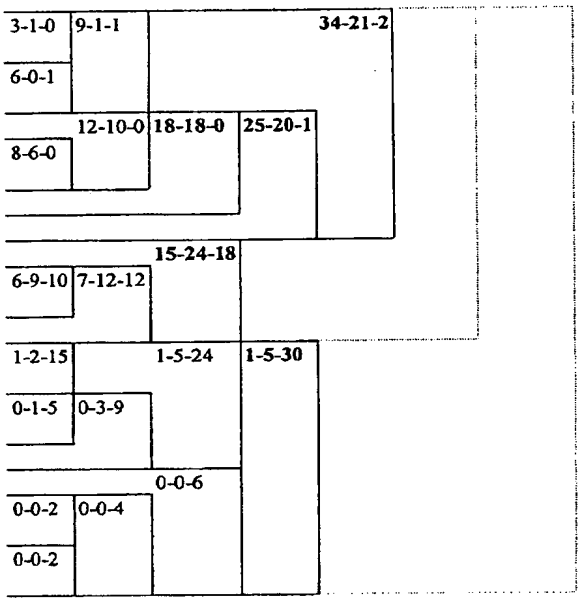


Tableau 7

Au bout d'une vingtaine d'étapes, on arrive à la loi discrétisée suivante:

Largeur de sépale	Iris versicolor	Iris virginica	Iris Setosa	Total
]- ∞ ; 2.95[	34	21	2	57
[2.95; 3.35[	15	24	18	57
[3.35; ∞ [	1	5	30	36
Total	50	50	50	150

Tableau 8

La valeur du χ^2 associée à la loi discrétisée est de 70,74, ce qui correspond à une probabilité d'indépendance de 1,66 10^{-14} (loi du χ^2 à 4 degrés de libertés). Deux fusions d'intervalles sont encore possibles. La meilleure d'entre elles est la première fusion, qui correspond à un χ^2 de valeur 54,17. La probabilité d'indépendance associée est 1,73 10^{-12} (loi du χ^2 à 2 degrés de libertés). Cette fusion ne respecte pas la condition (12) (elle augmente la probabilité d'indépendance) et est donc refusée.

L'attribut « largeur de sépale » a été discrétisé en 3 intervalles. Dans le premier intervalle, la classe Iris setosa est très rare. Dans les second, il y a équilibre entre les trois classes et dans le dernier, la classe Iris setosa est de loin la plus fréquente. Cette partition est celle qui minimise la probabilité d'indépendance des attributs « largeur de sépale » et « classe de la fleur ».

Nous envisagerons maintenant le cas où l'attribut à discrétiser est multi-dimensionnel, c'est-à-dire où l'attribut peut s'exprimer comme un vecteur $S=(S^1, ..,S^D)$ où D est la dimension de l'attribut et S^d, $d=1,..,D$ sont des attributs mono-dimensionnels. Nous considérerons pour simplifier le cas d'un attribut numérique bi-dimensionnel ($D=2$). Chaque individu peut alors être représenté comme un point ayant pour coordonnées les modalités de S^1 et S^2 de l'individu. La population des N individus de la base de donnée peut être ainsi « projetée » dans un plan (S^1, S^2) sous la forme d'un ensemble $\mathfrak{E}$ de points. Les relations de voisinage entre ces points peuvent être visualisée à partir du diagramme de Voronoï de l'ensemble $\mathfrak{E}$. On rappelle que le diagramme de Voronoï associé à un ensemble $\mathfrak{E}$ de points est une partition de l'espace (ici un plan) en cellules contenant chacune un point de $\mathfrak{E}$, chaque cellule étant définie comme l'ensemble des points de l'espace qui sont plus proches d'un point donné de $\mathfrak{E}$ que de tous les autres points de $\mathfrak{E}$. Une cellule est formée d'un polyèdre (ici un polygone) convexe entourant un point de $\mathfrak{E}$, chaque face du polyèdre étant un plan médiateur du point de $\mathfrak{E}$ associé à la cellule et d'un point voisin. A titre d'exemple, un diagramme de Voronoï associé à un ensemble de points est représenté en Fig. 4. A partir du diagramme de Voronoï on peut construire un diagramme dual, dit diagramme de Delaunay, reliant les points de $\mathfrak{E}$ appartenant à des cellules adjacentes. On a représenté en Fig. 5 le diagramme (ou graphe) de Delaunay associé au diagramme de Voronoï de la Fig. 4. Chaque arc du graphe de Delaunay représente une relation de voisinage entre deux points de $\mathfrak{E}$.

La méthode de discrétisation construit le graphe de Delaunay de $\mathcal{E}$ et utilise les arcs du graphe de Delaunay pour effectuer une partition de l'espace en zones élémentaires. Plus précisément, le graphe se compose d'arcs directs et d'arcs indirects. Les arcs directs entre deux noeuds ne passent que par les deux cellules adjacentes associées à ces noeuds. Le long d'un arc direct, le plus proche voisin est toujours un des deux points des deux cellules adjacentes. Les arcs indirects passent par au moins une troisième cellule de Voronoï. Le long d'un arc indirect, le plus proche voisin peut être un troisième point n'appartenant pas à une des deux cellules adjacentes. Lors d'un prétraitement, les arcs indirects sont éliminés. Seuls les arcs directs, traduisant une relation directe de proximité sont pris en compte lors de l'initialisation de la méthode de discrétisation. La fusion des cellules de Voronoï selon les arcs directs du graphe de Delaunay fournit les zones élémentaires.

Après avoir effectué une partition de l'espace en zones élémentaires, la méthode de discrétisation opère itérativement par fusion de zones, les seules fusions autorisées étant indiquées par un arc (direct) dans le graphe de Delaunay. Comme dans le cas mono-dimensionnel la fusion de deux zones n'est réalisée que si la condition (12) est vérifiée, c'est-à-dire que si cette fusion conduit à une diminution de la probabilité d'indépendance des attributs S et T. La discrétisation fournit des régions connexes, chaque région étant en fait une réunion connexe de cellules de Voronoï. Chaque région regroupe des individus homogènes statistiquement vis à vis de l'attribut cible et *a contrario* deux régions distinctes ont un comportement distinct vis à vis de cet attribut.

En outre, comme pour le cas mono-dimensionnel la valeur de probabilité d'indépendance obtenue à l'issue de la discrétisation permet de comparer les paires (de manière générales les n-uplets) d'attributs continus et de les classer en fonction de leur valeur prédictive d'un attribut cible.

La méthode de discrétisation multi-dimensionnelle s'applique encore à un attribut symbolique multi-dimensionnel, c'est-à-dire à un attribut $S=(S^1, ..,S^D)$ où S^d sont des attributs symboliques. Comme dans le cas mono-dimensionnel on construit un graphe dont les nœuds sont des modalités ou des groupes de modalités et l'on spécifie par des arcs les fusions possibles entre groupes.

A titre d'exemple, la Fig. 6 représente une population d'individus d'une base de données projetée sur le plan défini par deux attributs numériques continus. L'attribut

cible est la classe des individus pouvant prendre la modalité « classe 1 » représentée par un losange ou la modalité « classe 2 » représentée par un point.

La Fig. 7 représente le diagramme de Delaunay associé. On rappelle que l'on ne retiendra de ce diagramme que les arcs directs pour initialiser la liste des fusions possibles.

La méthode de discrétisation telle qu'exposée ci-dessus conduit à quatre zones, indiquées en Fig. 8 par des niveaux de gris différents. Ces zones connexes sont formées par la fusion de cellules de Voronoï contenant chacune un individu de la population initiale. La discrétisation permet de visualiser le comportement du couple d'attributs numérique vis à vis de l'attribut cible. Dans l'exemple représenté, on observera une relation de dépendance en spirale entre le couple d'attributs et l'attribut cible. Le tableau de contingence est en fait le suivant :

	Classe 1	Classe 2	Effectifs
Zone 1	11,8%	88,2%	212
Zone 2	2,5%	97,5%	122
Zone 3	88,7%	11,3%	512
Zone 4	69,5%	30,5%	154

Tableau 9

Ainsi, les zones 1 et 2 sont très majoritairement constituées d'individus de la classe 2 alors que la zone 3 est essentiellement constituée d'individus de la classe 1.

REVENDICATIONS

1) Méthode de discrétisation d'un attribut d'une base de données contenant une population d'individus, ledit attribut, dit attribut source, pouvant prendre plusieurs modalités, caractérisée en ce que, dans une première étape, on regroupe lesdites modalités de l'attribut source en groupes élémentaires et, qu'à partir du tableau de contingence de l'attribut source et d'un attribut cible, on détermine, dans une seconde étape, parmi un ensemble de couples de groupes élémentaires, le couple de groupes élémentaires dont la fusion diminue le plus fortement la probabilité d'indépendance de l'attribut source et de l'attribut cible, et que l'on fusionne dans une troisième étape le couple de groupes élémentaires ainsi déterminé, lesdites seconde et troisième étapes étant itérées tant qu'il existe un couple de groupes élémentaires permettant de diminuer ladite probabilité d'indépendance.

2) Méthode de discrétisation selon la revendication 1, caractérisée en ce que, pour déterminer le couple de groupes élémentaires dans la seconde étape, on estime pour chaque couple de groupes élémentaires dudit ensemble, la valeur du χ^2 du tableau de contingence après fusion dudit couple et l'on sélectionne le couple produisant la valeur du χ^2 après fusion la plus élevée.

3) Méthode de discrétisation selon la revendication 2, caractérisée en ce que, pour chaque couple de groupes élémentaires, on calcule la variation du χ^2 du tableau de contingence avant et après fusion dudit couple.

4) Méthode de discrétisation selon la revendication 3, caractérisée en ce que les variations du χ^2 associées aux différents couples sont triées sous forme de liste de valeurs décroissantes et que l'on sélectionne le premier couple de la liste.

5) Méthode de discrétisation selon l'une des revendications 2 à 4, caractérisée en ce que, le couple de groupes élémentaires étant sélectionné, on procède à la fusion dudit couple si la probabilité du χ^2 relative au tableau de contingence après fusion dudit couple est inférieure à la probabilité du χ^2 relative au tableau de contingence avant fusion.

6) Méthode de discrétisation selon la revendication 5, caractérisée en ce que les probabilités du χ^2 relatives au tableau de contingence avant et après fusion sont exprimées de manière logarithmique.

7) Méthode de discrétisation selon l'une des revendications précédentes, carcatérisée en ce que ledit ensemble de couples de groupes élémentaires est constitué de tous les couples de groupes voisins au sens d'une relation de voisinage prédéterminée.

8) Méthode de discrétisation selon la revendication 7, caractérisée en ce que l'on recherche parmi les couples de groupes élémentaires voisins ceux comprenant au moins un groupe présentant au moins un effectif théorique par case du tableau de contingence inférieur à un effectif minimum prédéterminé et qu'on les identifie comme couples prioritaires au moyen d'une information d'identification.

9) Méthode de discrétisation selon la revendication 8, caractérisée en ce que, s'il existe un ou des couples prioritaires, on fusionne le couple prioritaire produisant la valeur du χ^2 après fusion la plus élevée.

10) Méthode de discrétisation selon l'une des revendications 7 à 10, caractérisée en ce que, l'attribut source étant un attribut numérique mono-dimensionnel, les groupes élémentaires voisins sont constitués par des intervalles adjacents.

11) Méthode de discrétisation selon l'une des revendications 7 à 10, caractérisée en ce que, l'attribut source étant un attribut numérique multi-dimensionnel formé par pluralité d'attributs numériques mono-dimensionnels et les individus de la population étant représentés par des points dans l'espace desdits attributs, lesdits groupes élémentaires sont les cellules de Voronoï de cet espace, contenant lesdits points.

12) Méthode de discrétisation selon la revendication 11, caractérisée en ce que l'on construit le graphe de Delaunay associé aux cellules de Voronoï et que l'on élimine de ce graphe tout arc joignant deux cellules voisines en passant par une troisième, les couples de groupes élémentaires voisins étant alors donnés par les arcs du graphe de Delaunay après l'étape d'élimination.

13) Méthode de discrétisation selon l'une des revendications 7 à 10, caractérisée en ce que l'attribut source est de type symbolique.

14) Méthode d'évaluation de la dépendance d'un attribut d'une base de données vis à vis d'un attribut cible, caractérisée en ce que ledit attribut est discrétisé par la méthode de discrétisation selon l'une des revendications 1 à 13 et que la dépendance dudit attribut est estimée à partir de la probabilité de la valeur du χ^2 de l'attribut ainsi discrétisé.

15) Méthode d'évaluation de la dépendance d'un attribut numérique bi-dimensionnel, formé par un couple d'attributs numériques mono-dimensionnels, vis à vis d'un attribut cible et les individus de la population étant représentés par des points dans le plan desdits attributs, caractérisée en ce que l'attribut bi-dimensionnel est discrétisé par la méthode de discrétisation selon la revendication 12 et que l'on visualise par des moyens de visualisation des goupes de cellules de Voronoï fusionnées par ladite méthode.

16) Logiciel de data mining comprenant un programme de discrétisation d'au moins un attribut d'une base de données, caractérisé en que son exécution sur un ordinateur effectue les étapes de la méthode revendiquée selon l'une des revendications précédentes.

2825168

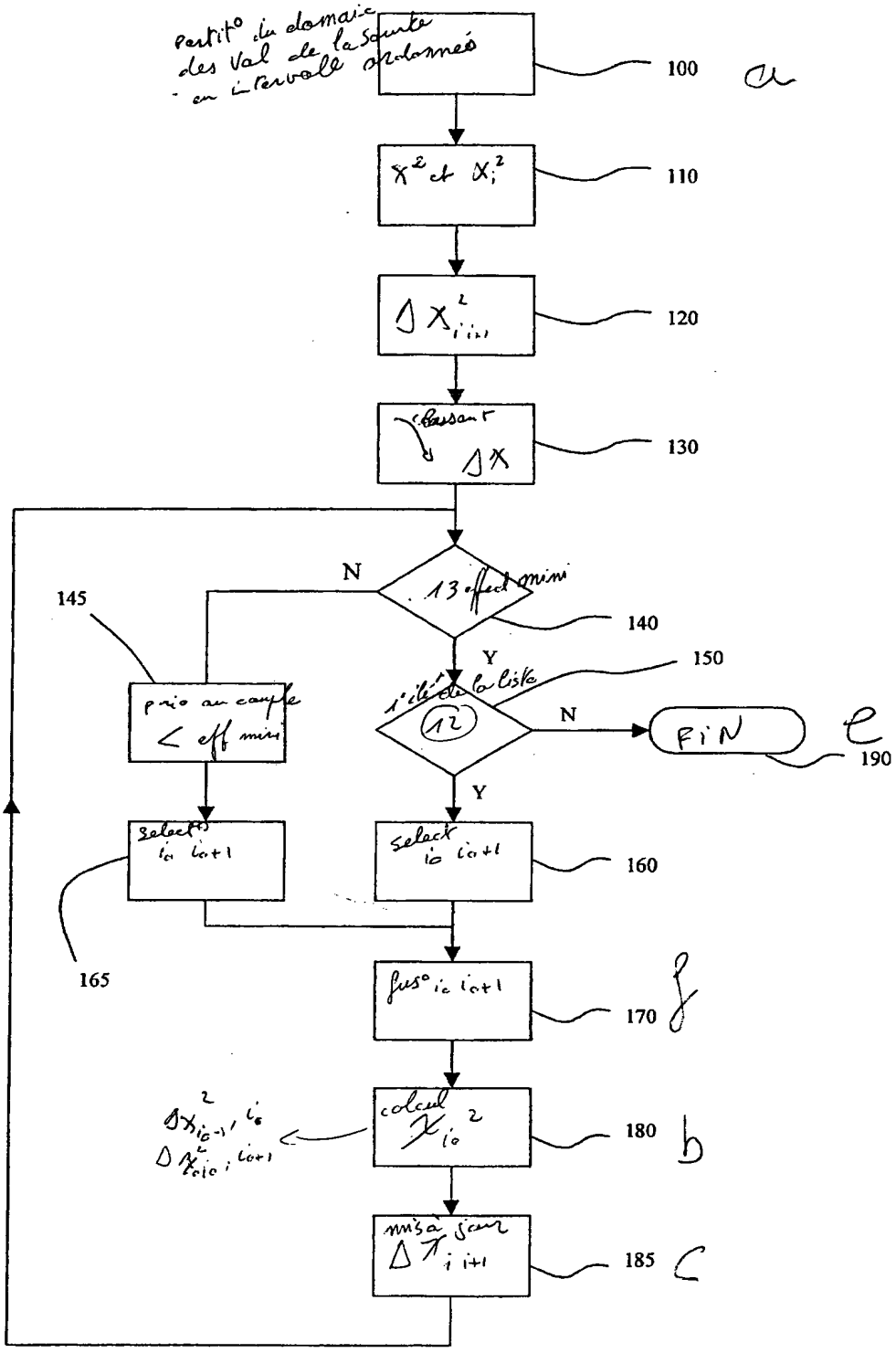


Fig. 1

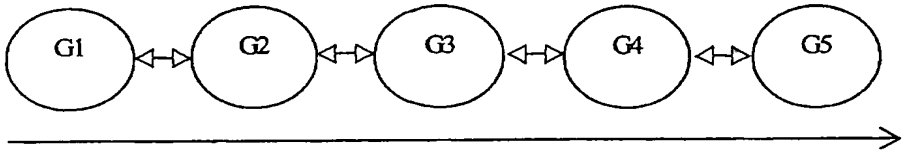


Fig. 2

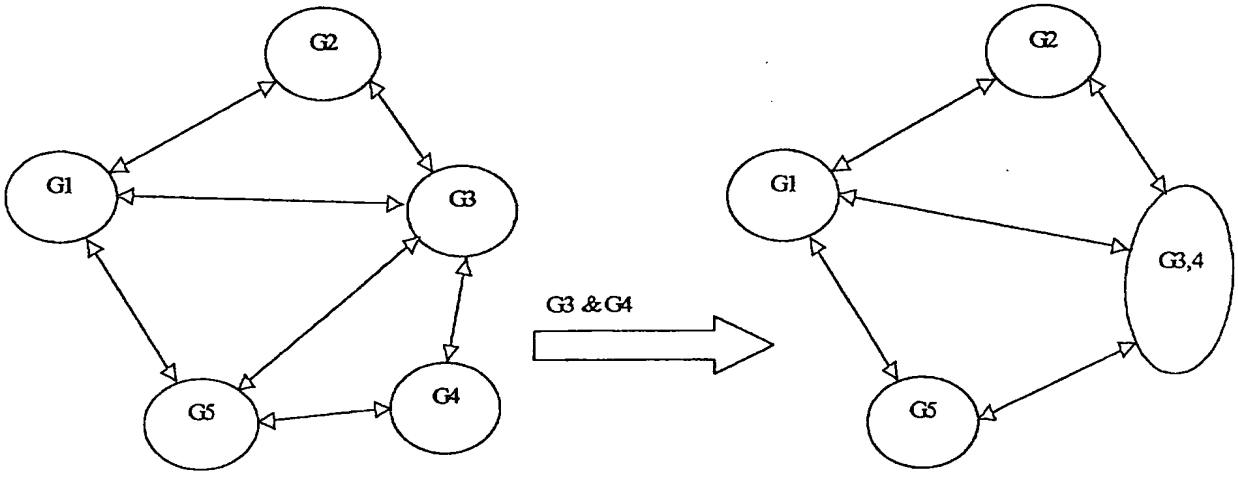


Fig. 3

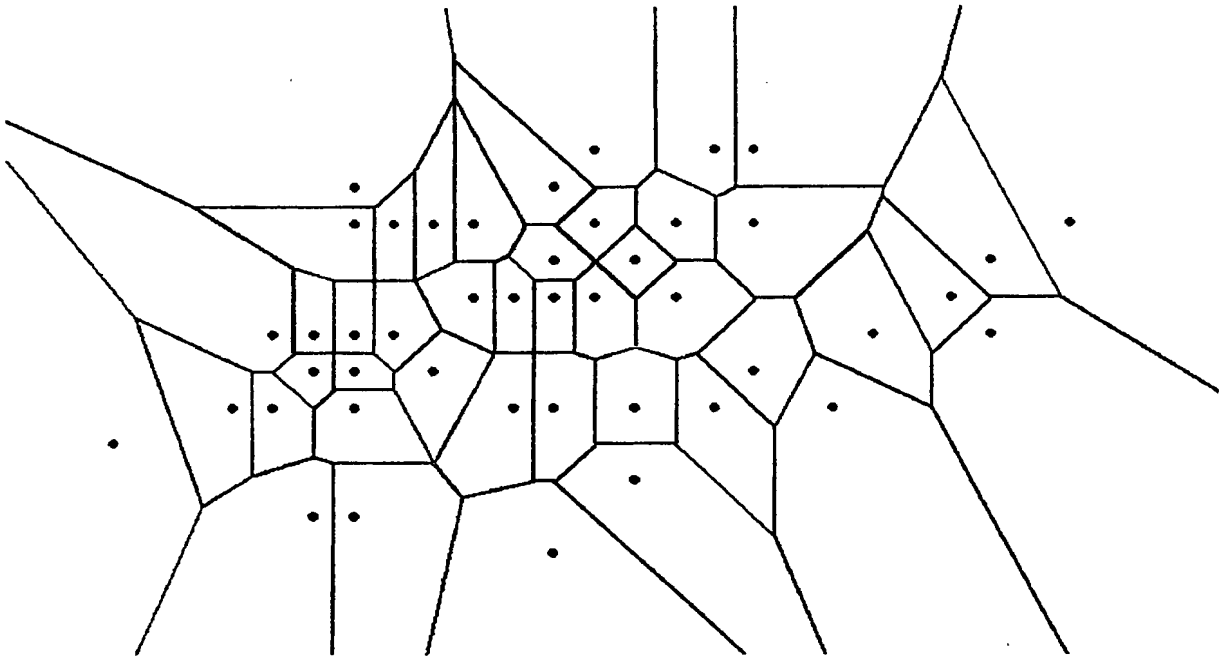


Fig. 4

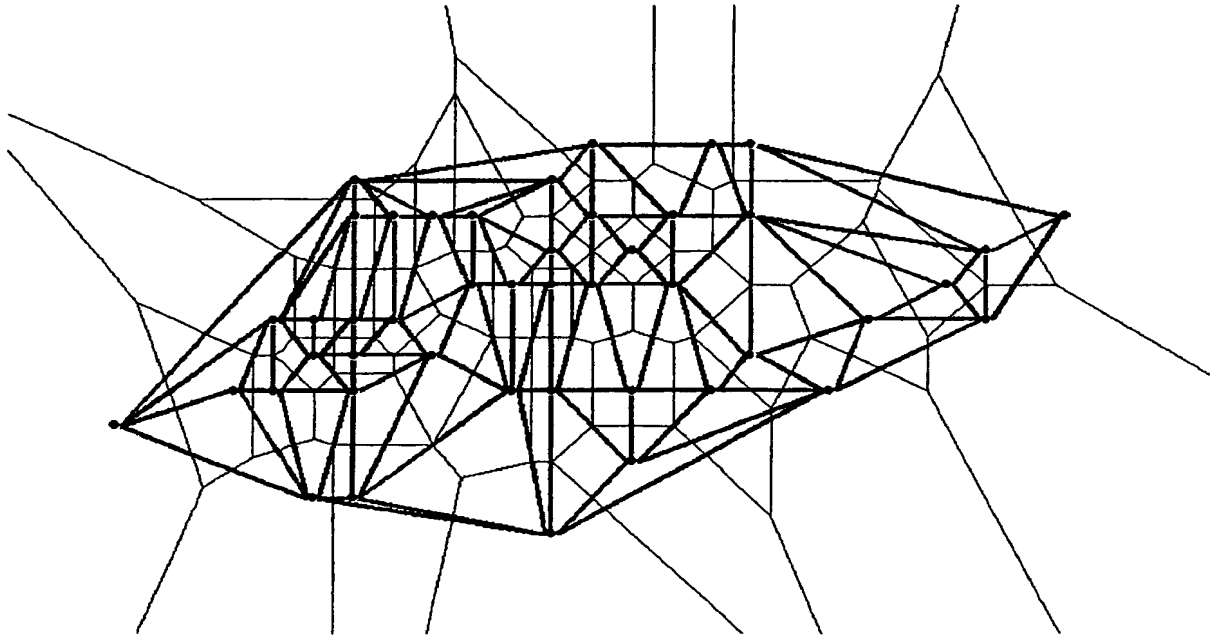


Fig. 5

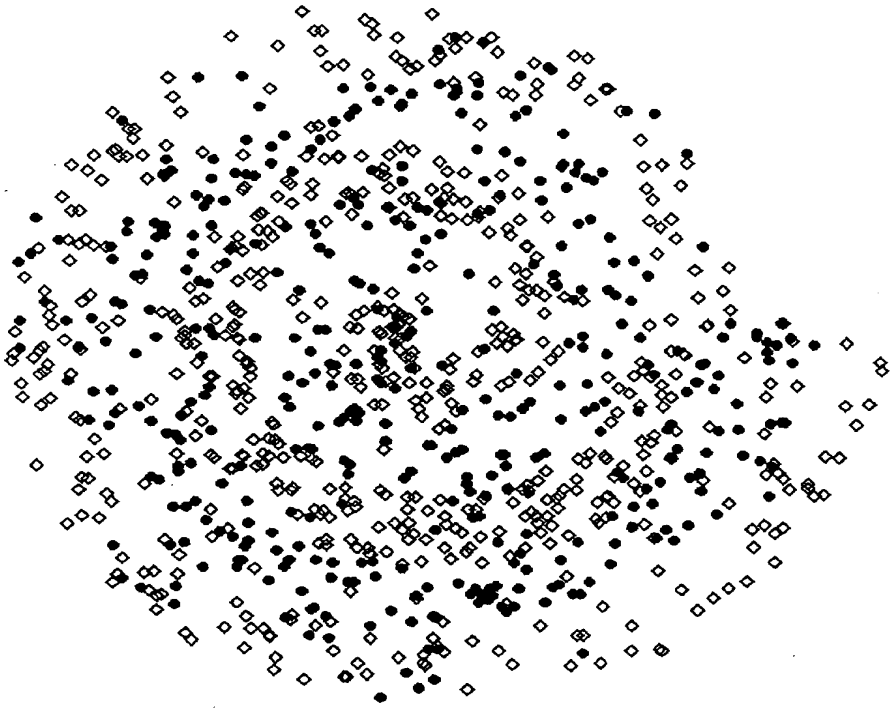


Fig. 6

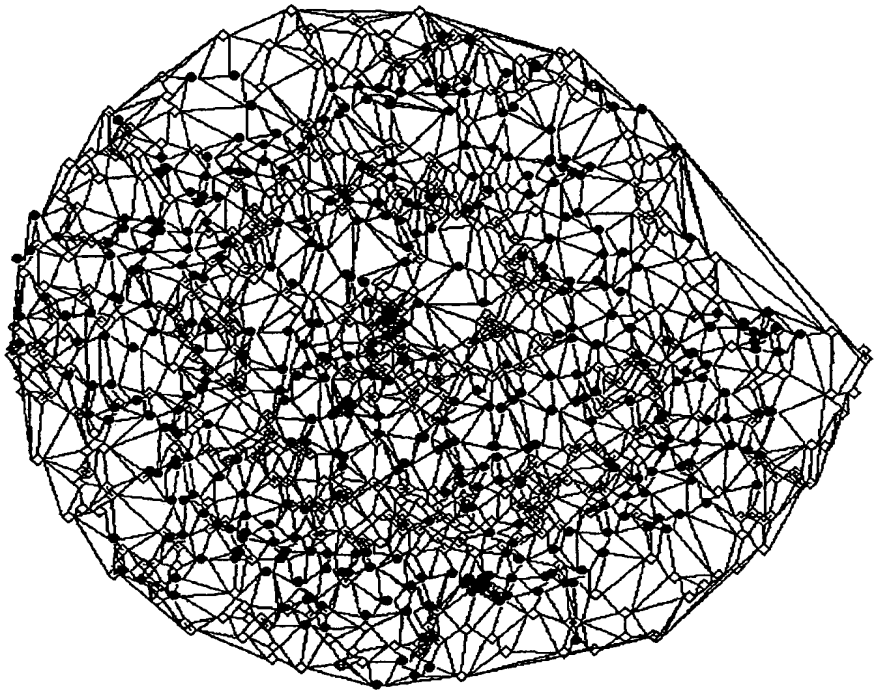


Fig. 7

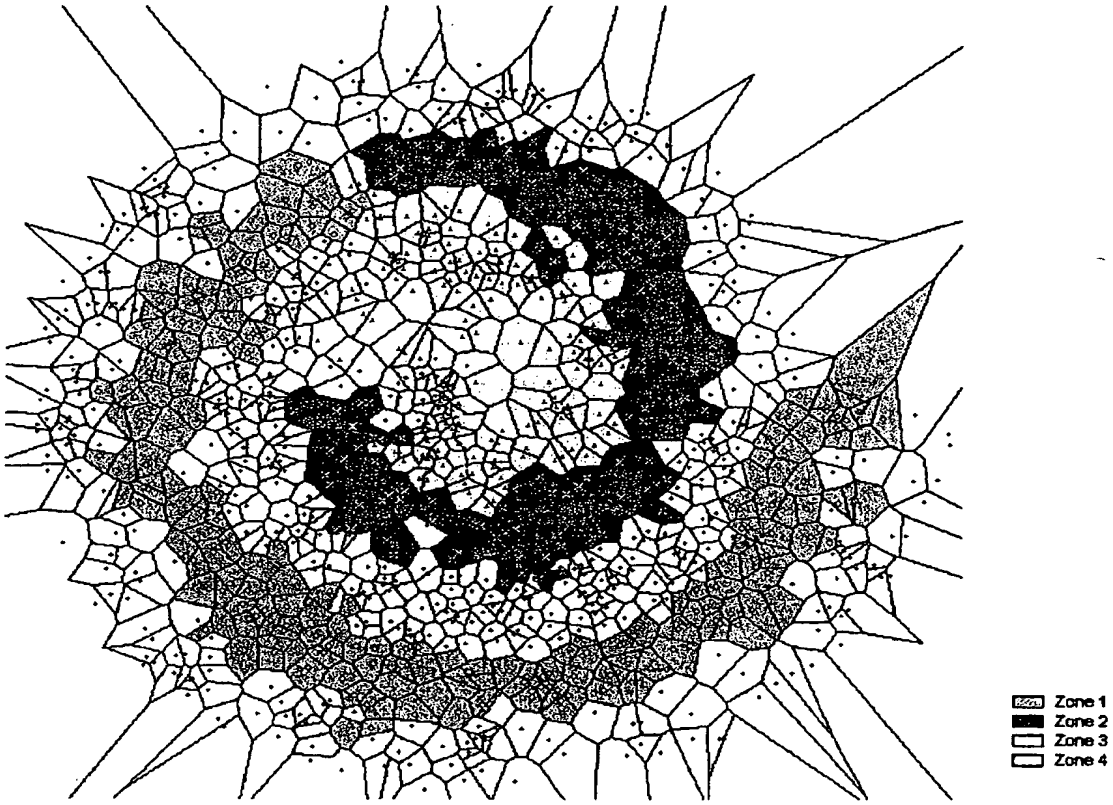


Fig. 8

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

2825168

N° d'enregistrement national

FA 615077
FR 0107006

RAPPORT DE RECHERCHE PRÉLIMINAIRE

établi sur la base des dernières revendications déposées avant le commencement de la recherche

DOCUMENTS CONSIDÉRÉS COMME PERTINENTS		Revendication(s) concernée(s)	Classement attribué à l'invention par l'INPI
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes		
A	HUAN LIU ET AL: "Chi2: feature selection and discretization of numeric attributes" PROCEEDINGS. SEVENTH INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE (CAT. NO.95CB35878), PROCEEDINGS OF 7TH IEEE INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE, HERNDON, VA, USA, 5-8 NOV. 1995, pages 388-391, XP002204465 1995, Los Alamitos, CA, USA, IEEE Comput. Soc. Press, USA ISBN: 0-8186-7312-5 * page 388, colonne de droite, alinéa 3 - page 389, colonne de droite, alinéa 1 * ---	1-16	G06F17/10
A	KERBER R: "ChiMerge: discretization of numeric attributes" AAAI-92. PROCEEDINGS TENTH NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, SAN JOSE, CA, USA, 12-16 JULY 1992, pages 123-128, XP002204466 1992, Menlo Park, CA, USA, AAAI Press, USA * page 124, colonne de droite, alinéa 3 - page 125, colonne de gauche, alinéa 1 * --- -/--	1-16	DOMAINES TECHNIQUES RECHERCHÉS (Int.CL.7) G06F G06K

1

Date d'achèvement de la recherche	Examinateur
3 juillet 2002	Correia Martins, F

CATÉGORIE DES DOCUMENTS CITÉS

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : arrière-plan technologique
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure.
D : cité dans la demande
L : cité pour d'autres raisons

& : membre de la même famille, document correspondant

EPO FORM 1503 12.99 (P04C14)

RÉPUBLIQUE FRANÇAISE

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

RAPPORT DE RECHERCHE PRÉLIMINAIRE

établi sur la base des dernières revendications déposées avant le commencement de la recherche

2825168
N° d'enregistrement national

FA 615077
FR 0107006

DOCUMENTS CONSIDÉRÉS COMME PERTINENTS		Revendication(s) concernée(s)	Classement attribué à l'invention par l'INPI
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes		
A	DOUGHERTY J ET AL: "Supervised and unsupervised discretization of continuous features" MACHINE LEARNING. PROCEEDINGS OF THE TWELFTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, PROCEEDINGS OF 12TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, TAHOE CITY, CA, USA, 9-12 JULY 1995, pages 194-202, XP002204467 1995, San Francisco, CA, USA, Morgan Kaufmann Publishers, USA * abrégé *	1-16	
A	BAY S D: "Multivariate discretization for set mining" KNOWLEDGE AND INFORMATION SYSTEMS, NOV. 2001, SPRINGER-VERLAG, UK, vol. 3, no. 4, pages 491-512, XP002204468 ISSN: 0219-1377 * page 3, alinéa 5 - page 5, alinéa 2 * * page 7, alinéas 2-8 * * page 8, alinéas 5-7 *	1-16	DOMAINES TECHNIQUES RECHERCHÉS (Int.CL.7)

Date d'achèvement de la recherche	Examinateur
3 juillet 2002	Correia Martins, F

CATÉGORIE DES DOCUMENTS CITÉS

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : arrière-plan technologique
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure.
D : cité dans la demande
L : cité pour d'autres raisons

& : membre de la même famille, document correspondant

EPO FORM 1503 12.99 (P04C14)